Journal of Systems Architecture 80 (2017) 85–91

Contents lists available at ScienceDirect

Journal of Systems Architecture

journal homepage: www.elsevier.com/locate/sysarc

Using data visualization technique to detect sensitive information re-identification problem of real open dataset☆

Chiun-How Kao[a,c], Chih-Hung Hsieh[*,b], Yu-Feng Chu[b], Yu-Ting Kuang[b], Chuan-Kai Yang[a]

[a] Department of Information Management, National Taiwan University of Science and Technology, Taipei, Taiwan
[b] Institute for Information Industry, Taipei, Taiwan
[c] Institute of Statistical, Academia Sinica, Taipei, Taiwan

ARTICLE INFO

Keywords:
Data de-identification
Data visualization
Privacy preserving
Personally identifiable information
Sensitive personal information

ABSTRACT

With plenty valuable information, open data are often deemed as great assets to academia or industry. In spite of some de-identification processing that most of data owners will perform before releasing the data, the more datasets are opened to public, the more likely personal privacy will be exposed. According to previous real case studies, even though the personally identifiable information has been de-identified, sensitive personal information could still be uncovered by heterogeneous or cross-domain data joining operations. The involved privacy re-identification processes are usually too complicated or obscure to be realized by data owners, not to mention that this problem will be more severe as the scale of data will get larger and larger. For preventing the leakage of sensitive information, this paper shows how to use a novel visualization analysis tool for open data de-identification (ODD Visualizer) to verify whether there exists sensitive information leakage problem in the target datasets. The high effectiveness that the ODD Visualizer can provide mainly comes from implementing a scalable computing platform as well as developing an efficient data visualization technique. Our demonstrations show that the ODD Visualizer can indeed uncover real vulnerability of record linkage attacks among open datasets available on the Internet.

1. Introduction

As everyone already knows, releasing dataset as public "open data" provides valuable information to academia or industry, and in fact it results in huge market size up to 3000 billion among various domains [1]. Despite the significant potential that open data can offer, Personally Identifiable Information (PII) re-identification or sensitive privacy leakage problem becomes an annoying side effect accompanied with the dataset releasing, and is one of the primary causes that only 10% amount of datasets owned by worldwide governments have been released [2,3]. Even with a dataset being de-identified, there still exists a chance that its PII may be revealed by cross joining different datasets [4,5]. For instance, L. Sweeny ever proposed some practical surveys claiming that (1) by correlating the National Association of Health Data Organizations (NAHDO) data and voter registration list for Cambridge Massachusetts via attributes of birthday, gender, ZIP code, six people had Governor Weld's particular birth date; only three of them were men; and he was the only one in his 5-digit ZIP code. Therefore, Weld's health PII can be subsequently identified [4]. and (2) 40% of 1130 volunteers who donated their DNA data to the Personal Genome Project can be identified, as well as using the zip code, date of birth, and gender [5]. Although their names did not appear, their profiles list sensitive medical conditions including abortions, illegal drug use, alcoholism, depression, sexually transmitted diseases, medications and their genome-related data.

To quantify the likelihood where PII or sensitive privacy may be re-identified, the *k*-anonymity model was proposed and can be used to measure how likely the PII of released data being de-identified [6] as the following:

k*-anonymity model**: for any given dataset table ***T, and one record ***D*** in ***T***, we assume that attributes = $\{V_1, V_2, \ldots\}$ contained in ***D*** can be partitioned into four groups according to their roles during the process of de-identification or re-identification; that is, ***D*** = (***Explicit Identifier, Quasi Identifier, Sensitive Attributes, Non-Sensitive Attributes***), where (1) ***Explicit Identifier*** (***EID***): is a set of attributes, such as name and social security number (SSN), containing information that explicitly identifies record owners; (2) ***Quasi Identifier*** (***QID***): is a set of attributes that could potentially identify record owners; (3) ***Sensitive Attributes***

☆ Fully documented templates are available in the elsarticle package on CTAN.

* Corresponding author.
E-mail addresses: maokao@stat.sinica.edu.tw (C.-H. Kao), chhsieh@iii.org.tw (C.-H. Hsieh), leon@iii.org.tw (Y.-F. Chu), ytkuang@iii.org.tw (Y.-T. Kuang), ckyang@cs.ntust.edu.tw (C.-K. Yang).

http://dx.doi.org/10.1016/j.sysarc.2017.09.009
Received 28 February 2017; Received in revised form 12 September 2017; Accepted 27 September 2017
Available online 28 September 2017
1383-7621/

(***SA***): consist of sensitive personal-specific information such as disease, salary, and disability status; and (4) ***Non-Sensitive Attributes*** (***NSA***): contain all attributes that do not fall into the above three categories. Assume ***qid*** is one existing value of one valid ***QID*** combination. For any ***qid*** of each ***QID*** in ***T***, if there are at least k records sharing the same ***qid***, then such ***T*** satisfying this requirement is called k-anonymous. The probability of linking a victim to a specific record through this ***QID*** is therefore at most $1/k$.

Time complexity of calculating the optimal k-anonymity value for a given dataset is known to be Non-deterministic Polynomial-time Hard (NP-hard) [7]. This means that when a dataset goes to an extremely large-scale, estimating whether it has high risk of privacy leakage could becomes intractable. In this paper, we show how to use a novel and efficiently scalable visualization tool, which is named as Open Data De-identification Visualizer (ODD Visualizer) [8], to estimate the likelihood or risk that the sensitive information in datasets will be re-identified. The core techniques in the ODD Visualizer, including Matrix Visualization (MV) approach [9] and Hierarchical Analysis and Clustering Tree (HACT), are used to depict a brief k-anonymity distribution among different attribute subsets as well as an optimal alignment of attributes, to provide a user the most robust attribute subset. In addition, we implemented the **ℓ-diversity model** to detect "Attribute Linkage Attack" individually. And the visualization of ℓ-diversity distribution shares the same interface as the one for k-anonymity distribution.

The merits of the proposed ODD Visualizer are threefold. (1) A scalable database and computation platform were incorporated in the ODD Visualizer such that k-anonymity of each attribute subset can be rapidly estimated. (2) Users can easily get a whole picture describing the k-anonymity and ℓ-diversity distribution among different attribute subset combinations. Hence, it can be known where the weakness of current dataset against PII re-identification is. (3) Based on the optimal alignment sorted by HACT, users get suggestion to decide which attribute subsets can be released or not. The efficiency and effectiveness were evaluated by one solid real case that uses the ODD Visualizer to detect a vulnerability of record linkage attack among real medical center data and census registration information.

Details about the definition of k-anonymity and ℓ-diversity model discussed in this paper can be found in Section 2. The architecture and implementation of the proposed ODD Visualizer are all mentioned in Section 3. Sections 4 is for demonstrating the effectiveness of our proposed method using a benchmark and real datasets available on the Internet. At last, Section 5 concludes this paper as well as gives some possible directions for further researches.

2. Related work

2.1. Record linkage attack and k-anonymity model

The key feature and major service of the ODD Visualizer is for information de-identification risk estimation and its visualization. The k-anonymity model [6] is adopted to detect the following "record linkage attack". The so-called "record linkage attack" and its relationship with k-anonymity model can be described using the following example. Suppose that a regional hospital is going to release patient's de-identified information (as regional patients' diagnosis table in Table 1(a)) for research purpose, and that there is another external table containing residents information of the same local region (as external table in Table 1(b)). Once a PII hacker has the privilege to access both public datasets, then the hacker does have a chance to identify the ***SA*** of "Disease" via ***QID*** = {job, gender, age}. For example, a privacy leakage crisis is on Doug, because he is the only one whose value of ***QID*** (named as ***qid***) is {lawyer, male, 38} (for the case of $k = 1$). Using such a ***qid*** value to link the two tables, the sensitive information of Doug's diagnosis result will leak out. On the contrary, PII for Bob and Fred is much safer, as they share the same *qid* of {engineer, male, 35} (for the case of $k > 1$). As a consequence, the k-anonymity model will be used to measure the likelihood of the sensitive information leakage. In the default scenario, a dataset owner can leverage the ODD Visualizer to detect the weakness of target a dataset.

Table 1
Examples for illustrating record linkage re-identification and k-anonymity model.

(a) Regional patients' diagnosis table

Job	Gender	Age	Disease
Engineer	Male	35	Hepatitis
Engineer	Male	38	Hepatitis
Lawyer	Male	38	HIV
Writer	Female	38	Flu
Writer	Female	30	HIV
Dancer	Female	30	HIV
Dancer	Female	30	HIV

(b) External residents table

Name	Job	Gender	Age
Alice	Writer	Female	30
Bob	Engineer	Male	35
Cathy	Writer	Female	30
Doug	Lawyer	Male	38
Emily	Dancer	Female	30
Fred	Engineer	Male	35
Glady	Dancer	Female	30
Henry	Lawyer	Male	39
Irene	Dancer	Female	32

2.2. Attribute linkage attack and ℓ-diversity model

Sometimes even k of the ***QID*** is bigger than one, an attacker can still re-identify the sensitive information from the correlation between ***QID*** and ***SA***. For example, all 30 years old female dancers have HIV, which is the sensitive information in Table 1(a), and there are just only two persons (***qid*** = {dancer, female, 30}) in Table 1(b). Therefore the attacker can infer that Glady and Emily are HIV patients with 100% confidence. The above situation is called "attribute linkage attack". To prevent attribute linkage attack, Machanavajjhala et al. [10] proposed a new model called ℓ-diversity model. The ℓ-diversity model makes use of the concept of *entropy* to estimate the diversity of the sensitive attribute in each *qid* group. The ℓ-diverse of every *qid* group is defined by the following equation:

$$log(\ell) = -\sum_{s \in S} P(qid, s) log(P(qid, s))$$

where S is one of the sensitive attributes and $P(qid, s)$ is the proportion of the sensitive value s in a *qid* group. The worst case is that all the values of the sensitive attributes in the *qid* group are the same, causing the entropy value of $\log(\ell)$ to be equal to 0, and it means that ℓ of the *qid* group is with the minimum value of 1. In our system we use the minimum of ℓ values of *qid* groups to define the ℓ of ***QID***. For example: there are 6 *qid* groups in Table 1(a), and the ℓ of (***qid*** = {dancer, female, 30} equals to 1, which is the minimum value, so the ℓ of ***QID*** is set to 1.

2.3. Privacy preserving data visualization

Recently, privacy preserving data visualization has attracted more attention in visualization community. In information visualization literature, Dasgupta et al. [11] described opportunities and challenges for privacy preserving visualization in the realm of electronic health record (EMRs) data. Andrienko et al. [12,13] discussed privacy issues in applying geospatial visual analytics methods to movement data. Dasgupta et al. [14,15] applied k-anonymity and ℓ-diversity in *parallel coordinate plot* and *scatter plots*. Chou et al. [16,17] integrated privacy issues in *Sankey diagram-like visualization* and *Node-link diagram*. In previous studies, they bundled a set of edges or merged a set of nodes to protect